36-732 Final Project: Examination of Optimal Treatment Regimes

Michael Kronovet (mkronove) [1]

Introduction

The study of optimal treatment regimes has been gaining traction recently especially as we enter the era of personalized medicine where medical treatments will be further tailored to the individual. Given a list of possible treatments, a **treatment regime** is list of rules defining which treatment should be selecting for which groups of people. For example, if our treatment was to give someone an antibiotic, some groups of people may be cured by the antibiotic, while others may suffer an allergic reaction if they receive it or not be impacted at all. A possible treatment regime in this case would be to assign the antibiotic to everyone under 40 years old and not assign it to everyone else, although this may not be the best one since age doesn't necessarily tell you who will have an allergic reaction. An **optimal treatment regime** is the treatment regime which gives us the best possible outcome. In our example, the optimal treatment regime could be the treatment rule that results in the largest number of people being cured. I will be reviewing recent literature on finding these optimal treatment regimes in this paper, with my primary focus being on **dynamic optimal treatment regimes**. Dynamic treatment regimes are treatment regimes where there are multiple time points where we get to decide treatment. Looking back at our example, if we got to re-examine patients every week to decide whether or not we should give them the antibiotic or not, then this would be a dynamic treatment regime. As you can see, the dynamic treatment regime generalizes the idea of the treatment regime that I introduced earlier (aka the static treatment regime). For simplicity, I will begin my analysis on optimal treatment regimes with static treatments where there is only a single treatment decision for each datapoint.

Static Treatment Regimes

Setup

Assume that treatment is binary and is denoted as A, where A takes on values of 0 or 1, for respectively not receiving or receiving treatment. X is a vector of subject characteristics that are measured before treatment is assigned. Y is the observed outcome of interest where we assume larger values of Y are preferred. The observed data (Y_i, A_i, X_i) is independent and identically distributed (iid) $\forall i = 1, ..., n$.

A treatment regime is a function $g : X \to \{0, 1\}$ that maps covariates to a treatment. $Y^*(1)$ and $Y^*(0)$ are the potential outcomes that would be observed if a subject received treatment $A = 1$ or $A = 0$, respectively. We assume that the observed outcome is the potential outcome for the treatment that was actually received, which can be expressed mathematically as $Y = Y^*(1)A + Y^*(0)(1 - A)$. We also assume that there are no unmeasured confounders which can be expressed as $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$. Therefore, under a specific treatment regime g, $Y(g) = Y^*(1)g(X) + Y^*(0)(1 - g(X))$.

Methods for Optimizing Static Treatment Regimes

A common approach for finding the optimal treatment regime is to pick the regime that maximizes the overall population mean so that $g^{opt} = \underset{g \in G}{\text{argmax}} \mathbb{E}[Y^*(g(X))]$, where G is the set of all treatment regimes. Under the assumptions that were outlined in the setup section, in this case we can say that $g^{opt}(X) = I\{\mathbb{E}[Y|A = 1, X] > \mathbb{E}[Y|A = 0, X]\}$, where the optimal regime assigns the treatment that yields the larger mean outcome conditional on the value of X.

In order to identify this optimal decision rule from the data, Zhang et al. (2012) suggests to use a parametric regression model for $\mathbb{E}[Y|A, X]$. However, Zhang et al. (2012) notes that only certain components of the parametric model of $\mathbb{E}[Y|A, X]$ will depend on the treatment. For example, if we had $\mathbb{E}[Y|A, X] = \beta_1 X_1 + \beta_2 A X_2$, then $\mathbb{E}[Y|A = 1, X] - \mathbb{E}[Y|A = 0, X]$ will only depend on β_2 since β_1 remains constant when the treatment changes. In this case we would only need to care about the term involving β_2 in our estimate of $\mathbb{E}[Y|A, X]$ in order to distinguish if a certain treatment had a better expected outcome. Now, suppose we only parameterize the useful subset of covariates whose impact on the expected outcome changes with treatment (X_2 in our example) using coefficients η. We can then find an optimal treatment regime that is parameterized by η, denoted as $g(X,)$, that is within the set of possible treatment regimes parameterized by η (aka G_η). Zhang et al. (2012) then proposes a doubly robust estimator for η which is consistent for $\mathbb{E}[Y^*(g_\eta)]$ if either $\pi(X; \gamma)$ (the parametric model estimate for the propensity score) or $\mu(A, X; \beta)$ (the parametric model estimate for $\mathbb{E}[Y|X, A]$),

but not both, is misspecified.

Zhao et al. (2012) takes a different approach where instead of attempting to optimize the potential outcomes, they try to maximize $\mathbb{E}_{g(X)}[Y]$. $\mathbb{E}_g[Y]$ denotes the expectation with respect to $\mathbb{P}_g$ which denotes the distribution (X, A, Y) given that $A = g(X)$. They also have a slightly different problem setup where $A = \{-1, 1\}$. Zhao et al. (2012) then shows that in this case$\mathbb{E}_{g(X)}[Y] = \mathbb{E}[\frac{I(A=g(X)}{A\pi+(1-A)/2}]$, where $\pi = P(A = 1)$, and that maximizing $\mathbb{E}_{g(X)}[Y]$ is equivalent to minimizing $\mathbb{E}[\frac{I(A\neq g(X)}{A\pi+(1-A)/2}]$. In order to transform this to a convex optimization function, this loss function is substituted for the following hinge loss $\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{A_i\pi+(1-A_i)/2}(1-A_i(f(X_i)))^+ + \lambda_n||f||^2$, where $x^+ = max(x, 0)$, $||f||$ is some norm for f, and $g(X) = sign(f(X))$. Support vector machines can then be used to solve for the optimal linear decision rules as well as more complex decision rules when using kernels. It is then proven that if $\bar{f}$ minimizes this hinge loss function, then the optimal treatment regime $\bar{g}(X) = sign(\bar{f}(X))$.

Dynamic Treatment Regimes

Murphy (2003) defines a dynamic treatment regime as a set of decision rules, with one rule for each time period. $\bar{a}_K = (a_1, ..., a_K)$ denotes the treatments that occurred at each time $t = 1, 2, ..., K$ and $\bar{S}_K = (S_1, ..., S_K)$ denotes the covariates that occur at each time. The jth decision rule will use information available from up to time j. The treatment decision made at time j is represented as a_j, and the vector of subject characteristics that are measured before treatment is assigned at the beginning of time interval by S_j.

Dynamic Treatment Regime Assumptions

According to Schulte et al. (2014), the following assumptions are standard when solving for the optimal treatment regimes:

- The consistency assumption that the covariates and outcomes observed in the study are those that potentially would be seen under the treatments actually received. This can be expressed as $S_k = S_k^*\left(\bar{A}_{k-1}\right), k = 2, \ldots, K$ and $Y = Y^*\left(\bar{A}_K\right)$.
- The stable unit treatment value assumption that assumes the covariates and outcome for a datapoint are unaffected by how treatments are allocated for any datapoint.
- The no unmeasured confounders assumption outlined earlier which can be expressed as $A_k \perp W^*|\bar{S}_k, \vec{A}_{k-1}$.

Data Sources

According to Schulte et al. (2014), data for dynamic treatment regimes typically comes from observational studies where participants are randomly sampled from the population and treatment assignment takes place according to routine clinical practice in the population. Data also often comes from intervention studies where treatments are randomized by those running the experiment. One such experimental design that has received a fair amount of attention to gather data for dynamic treatment regimes is the sequential multiple-assignment randomized trial (SMART). In a SMART, a participant receives a random treatment at each time point where the randomization probabilities for receiving a treatment a_k depend on $\bar{s}_k$ and $\bar{a}_{k-1}$. Since this takes the form of an experiment, the no unmeasured confounding assumption holds by design in a SMART, even though it is unverifiable in an observational study.

Q Learning

Typically, decision rules that maximize the mean response can be found using dynamic programming where the optimal rules are $d_1^*, \ldots, d_K^*$. This is done using the following algorithm:

Set
$J_K\left(\bar{S}_K, \bar{a}_{K-1}\right) = \sup_{a_K}\left(E\left[Y|\bar{S}_K, \bar{a}_{K-1}, a_K\right]\right)$
$d_K^*\left(\bar{S}_K, \bar{a}_{K-1}\right) = \arg\sup_{a_K}\left(E\left[Y|\bar{S}_K, \bar{a}_{K-1}, a_K\right]\right)$.

Then for each j, calculate

$$J_k\left(\bar{S}_k, \bar{a}_{k-1}\right) = \sup_{a_k}\left\{E\left[J_{k+1}\left(\bar{S}_{k+1}, \bar{a}_k\right)|\bar{S}_k, \bar{a}_{k-1}, a_k\right]\right\}$$

$$d_k^*\left(\bar{S}_k, \bar{a}_{k-1}\right) = \arg\sup_{a_k}\left\{E\left[J_{k+1}\left(\bar{S}_{k+1}, \bar{a}_k\right)|\bar{S}_k, \bar{a}_{k-1}, a_k\right]\right\}$$

Often, people estimate these expected values using regressions and refer to them as Q-functions. Q-learning, aka "quality learning", relies on using regression models on the outcome for the given relevant covariates at each decision point.

If we follow the convention of

$$Q_K\left(\bar{S}_K, \bar{A}_{K-1}, a_K\right) = E\left[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = a_K\right]$$

and

$$Q_k\left(\bar{S}_k, \bar{A}_{k-1}, a_k\right) = E\left[J_{k+1}\left(\bar{S}_{k+1}, \bar{A}_k\right)|\bar{S}_k, \bar{A}_{k-1}, A_k = a_k\right]$$

then

$$J_K\left(\bar{S}_K, \bar{A}_{K-1}\right) = \sup_{a_K : p_K\left(a_K | \bar{S}_K, \bar{A}_{K-1}\right) > 0} \left\{Q_K\left(\bar{S}_K, \bar{A}_{K-1}, a_K\right)\right\}$$

and

$$J_k\left(\bar{S}_k, \bar{A}_{k-1}\right) = \sup_{a_k : p_k\left(a_k | \bar{S}_k, \bar{A}_{k-1}\right) > 0} \left\{Q_k\left(\bar{S}_k, \bar{A}_{k-1}, a_k\right)\right\}$$

These Q-functions can be thought of as measuring the "quality" associated with using treatment a_k at decision k given the history up to that decision and then following the optimal regime thereafter. The functions J can be thought of as showing the "value" of a specific datapoint's history $\bar{a}_{k-1}$, $\bar{s}_k$ given that they received the optimal treatments in the future. Using Q-learning, we are able to find the optimal treatment regimes by directly modeling the Q-functions. However, if the Q-functions are not correctly specified then the estimated optimal treatment regime may not be a consistent estimator of the true optimal treatment regime.

Murphy (2003) proposes to instead minimize so-called regret functions via dynamic programming instead of directly the Q-functions. The regret function $\mu_k\left(\vec{S}_k, \vec{A}_{k-1}, a_k\right)$ can be thought of as the amount that we lose out on for making a sub-optimal decision a_k rather than the optimal decision at time k. The regret functions are defined by the following:

$$\mu_k\left(\bar{S}_k, \bar{A}_{k-1}, a_j\right) = J_k\left(\bar{S}_k, \bar{A}_{k-1}\right) - Q_k\left(\bar{S}_k, \bar{A}_{k-1}, a_k\right)$$

Murphy (2003) proposes to directly model this regret function using parametric, semiparametric, or nonparametric techniques (in the paper the author specifically uses parametric link functions to express the regret function). Given these regret functions we can redefine our conditional mean as follows and iteratively calculate the optimal treatments with dynamic programming:

$E\left[Y | \bar{S}_K, \bar{A}_K\right] = \mu_0 + \sum_{k=1}^{K} \phi_k\left(\bar{S}_k, \bar{A}_{k-1}\right) - \sum_{k=1}^{K} \mu_k\left(\bar{S}_k, \bar{A}_k\right)$

where $\mu_0 = E\left[J_1\left(S_1\right)\right]$ and $\phi_k\left(\vec{S}_k, \vec{A}_{k-1}\right) = J_k\left(\vec{S}_k, \vec{A}_{k-1}\right) - Q_{k-1}\left(\vec{S}_{k-1}, \vec{A}_{k-1}\right)$ for $k = 1, \ldots, K$.

This approach is an analogous to another popular approach for finding the optimal treatment regime known as A-learning.

A Learning

The main idea of A-learning, aka "advantage learning", is to propose alternative functions to the Q-functions like how in the above example Murphy (2003) estimated regret functions. The most common choice for the estimating function is the contrast in Q-functions between treatments with the same history of covariates/treatments. Therefore, A-learning relies on regression models for the outcomes from contrasting treatments as well as for the probability of observed treatment assignment given relevant covariates at each decision point.

We denote the contrast in Q functions for $k = 1, \ldots, K$ as $C_k\left(\bar{s}_k, \bar{a}_{k-1}\right) = Q_k\left(\bar{s}_k, \bar{a}_{k-1}, 1\right) - Q_k\left(\bar{s}_k, \bar{a}_{k-1}, 0\right)$. It is useful to note that $Q_k\left(\bar{s}_k, \bar{a}_k\right)$ may be written as $h_k\left(\bar{s}_k, \bar{a}_{k-1}\right) + a_k C_k\left(\bar{s}_k, \bar{a}_{k-1}\right)$, where $h_k\left(\bar{s}_k, \bar{a}_{k-1}\right) = Q_k\left(\bar{s}_k, \bar{a}_{k-1}, 0\right)$. This indicates that having $a_k = I\left\{C_k\left(\bar{s}_k, \bar{a}_{k-1}\right) > 0\right\}$ will maximize $Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k)$ where the maximum is $h_k\left(\bar{s}_k, \bar{a}_{k-1}\right) + C_k\left(\bar{s}_k, \bar{a}_{k-1}\right) I\left\{C_k\left(\bar{s}_k, \bar{a}_{k-1}\right) > 0\right\}$. In A-learning we typically estimate $C_k\left(\bar{s}_k, \bar{a}_{k-1}\right)$ and $h_k(\bar{s}_k, \bar{a}_{k-1})$ using some parametric or nonparametric techniques, which then allows us to identify the optimal treatment regime. After estimating these functions we can get our optimal treatment for the last time point $\hat{d}_K^* = I\left\{\hat{C}_K\left(\bar{s}_K, \bar{a}_{K-1}\right) > 0\right\}$, and we can find the optimal treatments from previous timepoints with dynamic programming.

Schulte et al. (2014) describes A-learning as a middle ground between Q learning and complex methods that employ flexible models for the Q-functions that are often difficult to interpret. This is due to the fact that A-learning allows us flexibly model the functions $h_k\left(\bar{s}_k, \bar{a}_{k-1}\right)$ while maintaining simple parametric models for the contrast functions $C_k\left(\bar{s}_k, \bar{a}_{k-1}\right)$. Since the treatment rule only depends on the parametric contrast function, the rule will be interpretable, while the model for the response can be more complex. Moreover, in A-learning identifying the optimal regime depends only on correct specification of the contrast or regret functions. This means that A-learning methods are is less sensitive to model misspecification than Q-learning. Although according to Chakraborty et al. (2010), Q-learning and A-learning can lead to the same estimators for the Q-function under certain conditions. One set of conditions is if the propensities scores for treatments are constant and linear models are used for the contrast and quality functions. When we extend the treatment to have more than two options at every stage, then A-learning tends to increase in complexity more than Q-learning (Schulte et al., 2014).

Optimal "Midstream" Treatment Regime

In real medical situations it is often the case that a new patient will be encountered after they have already received (or not received) some treatments that possibly don't follow the optimal regime. Suppose a new patient has already received treatments for the first $\ell-1$ treatment decision points where there are K treatment decision points total. How can we generalize our procedure to find the optimal treatment regime for a patient where we assign treatments from a regime starting at ℓ? Zhang et al. (2012) offers the solution. It is the case that the patient has a history due to the past treatments that can be viewed as realizations of the random variables $\left(S_1^{(P)}, A_1^{(P)}, \ldots, S_{\ell-1}^{(P)}, A_{\ell-1}^{(P)}, S_\ell^{(P)}\right)$. $A_k^{(P)}, k = 1, \ldots, \ell-1$, represent the treatments that the patient has already received prior to starting the optimal regime, and $S_k^{(P)}, k = 1, \ldots, \ell-1$, denote covariate information collected up to the ℓth decision. In order to get the new optimal treatment regime for this patient, we need to find the decision rules $d_k^{(t)}\left(\bar{s}_k, \bar{a}_{k-1}\right), \quad k = \ell, \ell+1, \ldots, K$, that optimize the response for a patient with a realized past history of $\left(\bar{S}_\ell^{(P)}, \bar{A}_{\ell-1}^{(P)}\right) = \left(\bar{s}_\ell, \bar{a}_{\ell-1}\right)$.

Then it is the case that for $k = K-1, \ldots, \ell$:

$$d_k^{(\ell)\text{opt}}\left(\bar{s}_k, \bar{a}_{k-1}\right) = \arg\sup_{a_k} \mathrm{E}\left[V_{k+1}^{(\ell)}\left\{\bar{s}_k \;\; S_{k+1}^*\left(\bar{a}_{k-1}, a_k\right), \bar{a}_{k-1}, a_k\right\} | \mathcal{V}_{\ell,k}\right]$$

$$V_k^{(\ell)}\left(\bar{s}_k, \bar{a}_{k-1}\right) = \sup_{a_k} \mathrm{E}\left[V_{k+1}^{(\ell)}\left\{\bar{s}_k \;\; S_{k+1}^*\left(\bar{a}_{k-1}, a_k\right), \bar{a}_{k-1}, a_k\right\} | \mathcal{V}_{\ell,k}\right]$$

Therefore, to optimize these functions and find the optimal treatment regime for a new patient that has a history up to point ℓ, we can use existing methods. It is important to note that the ℓth to Kth rules of the optimal regime for a patient with no prior history, $d^{(1)\text{ opt}}$, are not necessarily the same as the optimal rules for a new patient with a history up to ℓ, $d^{(\ell)\text{ opt}}$.

Direct Optimization Methods for Variable Selection

In 2015, Zhang & Zhang (2015) first introduced a direct optimization approach to attenuate the risk of model misspecification in commonly used DTR methods. Zhang & Zhang (2016) outlines that Q-learning and A-learning can be classified as outcome regression-based methods where the aim is to build good (parametric or semiparametric) regression models for outcomes given covariates. In these outcome regression-based approaches, the optimal treatment regimes are estimated by inverting the relationship between covariates and the regression models after the models are specified. It is clear that these methods are highly reliant on the assumption that the posited regression models are correctly specified. The method presented in Zhang & Zhang (2015) on the other hand, aims to directly maximize estimates $E\left\{Y^*(d)\right\}$ across a class of regimes.

They suggest using the (augmented) inverse probability weighted estimator (AIPWE)

$$\widehat{C}_{AIPWE}\left(X_i\right) = \frac{A_i}{\widehat{\pi}_i} Y_i - \frac{A_i - \widehat{\pi}_i}{\widehat{\pi}_i}\widehat{\mu}\left(1, X_i\right) - \left\{\frac{1-A_i}{1-\widehat{\pi}_i} Y_i - \frac{\widehat{\pi}_i - A_i}{1-\widehat{\pi}_i}\widehat{\mu}\left(0, X_i\right)\right\}$$

since it has the doubly robust property of being consistent if either treatment or outcome regression models, but not necessarily both, are correctly specified. Instead of simply using $I\{\hat{C}(X) > 0\}$ as their estimator for the optimal treatment regime, they then select the treatment regime that optimizes

$$d^{opt}(X) = \arg\min_{d \in \mathcal{D}} E[|C(X)| I\{Z \neq d(X)\}]$$

where $Z_i = I\left\{C\left(X_i\right) > 0\right\}$. Zhang & Zhang (2016) advocates that the additional step of optimization after obtaining $\hat{C}(X)$ is important for prescriptive variable selection and leads to the direct optimization $E\left\{Y^*(d)\right\}$, which is claimed to be more robust. From this method, Zhang & Zhang (2016) also proposes a direct optimization method for selecting variables that are useful for making treatment decisions, which I explain below.

Based on the optimization process in Zhang & Zhang (2015) and assuming a linear decision rule, the weighted misclassification error rate for a given regime $d(X)$ is

$$err(X_{j^1}, ..., X_{j^m}) = \frac{1}{n}\sum_{i=1}^{n}\left[|\hat{C}\left(X_i\right)| I\left\{\widehat{Z}_i \neq I(\beta_0 + \beta_1 X_{j^1} + ... + \beta_m X_{j^m} > 0)\right\}\right]$$

Given a set of selected prescriptive variables $\{X_{j^1}, ..., X_{j^m}\}$ and a potential prescriptive variable X_j, the difference in misclassification error

$$err(X_{j^1}, ..., X_{j^m}) - err(X_{j^1}, ..., X_{j^m}, X_j)$$

acts as a natural weighting of importance for the variable choice X_j. The prescriptive variable selection method is then as follows:

1. Iterate through all the covariates and calculate $err(null\ \ set) - err(X_j)$ (where we define $err(null\ \ set)$ as the above error where instead of selecting optimal betas we just pick the optimal choice between $I(\widehat{Z}_i \neq 0)$ and $I(\widehat{Z}_i \neq 1)$) and

include the k variables with the largest values of $err(null\ set) - err(X_j)$ in a set $\mathcal{F}^0$. Also add the covariate with the lowest $err(X_j)$ to the set $\mathcal{S}^{(1)}$ (denoted as $\mathcal{S}^{(m)}$ when having gone through m stages of the selection process), which will be the set of important prescriptive variables.

2. Keep adding new covariates to the set $\mathcal{S}$ where the m-th variable you add is $X_{j^m} = \arg\min_{X_j \in \mathcal{F}^0/\mathcal{S}^{(m-1)}} err(\mathcal{S}^{(m-1)}, X_j)$. Continue selecting features until the value of $\frac{err^{(m-1)} - err^{(m)}}{err^{(m-1)}}$ is below some threshold.

This selection process also holds in the case of dynamic treatments where you can identify the linear decision rule that minimizes the weighted misclassification error rate at each stage. All that must be changed is to convert the loss function to $\frac{1}{n}\sum_{i=1}^{n}\left[|\hat{C}_k\left(L_{ki}\right)|I\left\{\widehat{Z}_{ki} \neq d_k(L_i)\right\}\right]$ where $L_k = (\bar{X}_k, \bar{A}_k)$.

Experimentation

I estimated optimal dynamic treatment regimes using the BOWL estimator from Zhao et al. (2015), the AIWPE estimator from Zhang et al. (2012), and standard Q learning optimal treatment with the DynTxRegime R package. I applied these methods to a simulated dataset (provided in the DynTxRegime package) that mimics data from a two-stage randomized clinical trial studying the effect of meal replacement shakes on adolescent obesity. I tried to optimize for the negative percentage change of the BMI after 12 months from the original BMI which I will refer to as the "12 month percent change". Here are the results from each method:

BOWL At the first decision time point, I regressed the 4 month percent change on race, gender, parent BMI, and baseline BMI using a linear kernel for the regime. At the second decision time point, I regressed the 12 month percent change on race, gender, parent BMI, and BMI after 4 months using a linear kernel for the regime. I used a constant propensity score determined by a bernoulli glm regression. The BOWL method yielded negative coefficients for parent BMI, a positive coefficient for baseline BMI, and a negligible coefficient (very close to 0) for BMI after 4 months. The algorithm predicted that the optimal treatment regime would yield a mean decrease in weight over 12 months of 8.238923 percent. At decision point 1 the algorithm determined that 30.95238 percent of the patients should have received meal replacements and then 54.7619 percent at decision point 2.

AIPWE At the first decision time point, I regressed 12 month percent change on race, gender, parent BMI, and baseline BMI for my main effects model and 12 month percent change on race, gender, and baseline BMI for my outcome contrasts model. I used a constant propensity score determined by a bernoulli glm regression. At the second decision time point, I regressed 12 month percent change on race, gender, parent BMI, and BMI after 4 months for my main effects model and 12 month percent change on race, gender, and BMI after 4 months for my outcome contrasts model. The AIPWE method yielded negative coefficients for parent BMI, baseline BMI, and BMI after 4 months. This indicates that the higher your BMI measures are, the harder it is for you to lose a larger proportion of weight. The algorithm predicted that the optimal treatment regime would yield a mean decrease in weight over 12 months of 9.239559 percent. At decision point 1 the algorithm determined that 0.0 percent of the patients should have received meal replacements and then 37.14286 percent at decision point 2.

Q-learning I used the same model inputs as the AIPWE to implement the Q-learning method. Q-learning also yielded negative coefficients for parent BMI, baseline BMI, and BMI after 4 months. The algorithm predicted that the optimal treatment regime would yield a mean decrease in weight over 12 months of 7.196043 percent. At decision point 1 the algorithm determined that 4.285714 percent of the patients should have received meal replacements and then 65.7142 percent at decision point 2.

Each of these methods yielded similar predictions on the optimal decrease in weight over the 12 month study, so we could expect the best treatment sequence to have patients lose around 7 to 9.5 percent. The similarity in these results reassure that the methods are working correctly. However, there was a large amount of variety in the recommended treatments across the different methods, with the AIPWE method recommending that no one takes the meal replacement treatment at decision point 1 and the BOWL method estimating that around 30.9 percent should take the meal replacement treatment. This could indicate that there are many possible treatments combinations that yield large reductions in patient weight and each method is just discovering different combinations. It is difficult to say if the results these methods are accurate or not, since we do not have access to the responses under treatments we don't observe. The negative coefficients that both the Q learning and AIPWE methods deduced for covariates involving BMI did surprise me, because I thought that those with a higher BMI would need more food to maintain their weight. The positive coefficient for baseline BMI determined by the BOWL method makes more intuitive sense to me. In addition, it seems logical parent BMI would have a negative coefficient since patients who have parents with high BMI values may be genetically predisposed to having a large weight and would thus have a harder time losing weight. When I implemented each of these methods, none

of them specifically stood out. I had to rerun the BOWL method and the Q-learning method for every treatment decision point, which did not seem to be too efficient. Also, the AIPWE method seems to be yielding inaccurate results by claiming that no one in the first stage should take the treatment.

Conclusion

As you can see from this paper, there are a wide variety of techniques for computing the optimal treatment regimes. Most research in this field currently revolves around dynamic treatments, likely since dynamic methods are generalized and could be used to find the optimal static treatments. All the most popular methods for finding the dynamic treatment regimes are grounded in reinforcement learning that rely on dynamic programming. Q-learning, which is the standard reinforcement learning approach, is the most general method with most of the research being published today revolving around different adaptations of Q-learning. Many researchers who create new approaches based off Q-learning and A-learning techniques are dubbing their methods things such as C-learning (Zhang & Zhang, 2015) or V-learning (Luckett et al., 2019). It is apparent that research in the field is transitioning to approaches that do not rely as heavily on correct model specification. Techniques like ones outlined in Zhang & Zhang (2016) are also being adapted to not only learn optimal treatment regimes but to provide meaningful inference results.

Since dynamic treatment regimes has roots in reinforcement learning as well as causal inference, researchers with different backgrounds have been attracted the field. It seems the machine learning based approaches in Zhao et al. (2012) and Y.Q. et al. (2015) inspired the doubly robust estimator (which was more of a causal inference based approach) used in Zhang & Zhang (2015) and Zhang & Zhang (2016). There is much more research to come from this field, and it will be interesting to see how the approaches from the machine learning community and the causal inference community evolve and develop off of each other.

References

Chakraborty, B., Murphy, S., and Strecher, V. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343, 2010. doi: 10.1177/0962280209105013. URL https://doi.org/10.1177/0962280209105013. PMID: 19608604.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 0(0):1–34, 2019. doi: 10.1080/01621459.2018.1537919. URL https://doi.org/10.1080/01621459.2018.1537919.

Murphy, S. A. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B*, 65 (2):331–355, May 2003. doi: 10.1111/1467-9868.00389. URL https://ideas.repec.org/a/bla/jorssb/v65y2003i2p331-355.html.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. **Q**- and **A**-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.*, 29(4):640–661, 11 2014. doi: 10.1214/13-STS450. URL https://doi.org/10.1214/13-STS450.

Y.Q., Z., D., Z., E.B., L., and M.R., K. New statistical learning methods for estimating optimal dynamic treatment regimes. *Am. Stat. Assoc.*, 110: 583—598, 2015. doi: 10.1080/01621459.2014.937488. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517946/.

Zhang, B. and Zhang, M. C-learning: A new classification framework to estimate optimal dynamic treatment regimes: C-learning. *Biometrics*, 74, 12 2015. doi: 10.1111/biom.12836.

Zhang, B. and Zhang, M. Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariate. 2016.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics*, 68(4):1010–1018, December 2012. URL https://ideas.repec.org/a/bla/biomet/v68y2012i4p1010-1018.html.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. doi: 10.1080/01621459.2012.695674. URL https://doi.org/10.1080/01621459.2012.695674. PMID: 23630406.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583—598, 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.937488. URL https://europepmc.org/articles/PMC4517946.